Yu, G., & Zhang, J. (2017). Computer-Based English Language Testing in China: Present and Future. *Language Assessment Quarterly*, *14*(2), 177-188. https://doi.org/10.1080/15434303.2017.1303704

Computer-based English Language Testing in China: Present and Future

Abstract

In this special issue on high-stakes English language testing in China, the two articles on computer-based testing (Jin & Yan; He & Min) highlight a number of consistent, ongoing challenges and concerns in the development and implementation of the nation-wide IB-CET (Internet Based College English Test) and institutional computer-adaptive English tests, respectively: conceptualizing the construct of computer-based language testing, ensuring fairness for test takers with differing levels of computer literacy, and achieving comparability between tests or tasks of different delivery modes. In this article, we provide an overview of the research studies on computer-based English language testing conducted by Chinese scholars and published in major Chinese academic journals in recent decades, aiming to identify the research topics, gaps, and agendas that could have implications beyond Chinese contexts in promoting better use of computer technologies *in* and *for* English language testing.

Key words: Computer, China, language testing research

Special issues of international journals like *Language Assessment Quarterly* that focus on the regional development (e.g., Taiwan, Volume 9, Issue 1, 2012; mainland China, this issue; Japan, forthcoming) of English language testing are long overdue, given the increasing globalization and localization of high-stakes English language tests. Large-scale English language tests in China currently include international tests such as the Test of English as a Foreign Language, Internet-based Test (TOEFL iBT) and International English Language Testing System (IELTS), and numerous nationally or locally developed tests such as the College English Test (CET), Test for English Majors (TEM), Graduate School Entrance English Examination (GSEEE), Medical English Test System (METS) for nurses, National Matriculation English Test (NMET) for secondary school graduates, and Public English Testing System (PETS) for the general public (see Cheng, 2008 for brief introductions to the locally developed tests except for the METS). Among these locally developed tests, the CET is now Internet-based, and the speaking components of the PETS (Level 1, from September 2006) and NMET (in some provinces, e.g., Guangdong and Guangxi) are computer-based (Zeng, 2010). In addition, a few universities offer computer-based English tests to their students, often using commercially available platforms. Certain questions emerge in view of these developments: What is the current status of research on computer-based English language testing in China? To what extent are the current practices of computer-based English language testing in China supported by sufficient and strong research evidence from its unique social and educational contexts?

A systematic search that we conducted on the China Knowledge Resource Integrated Database[i]--using terms (all in Chinese) such as "computer/Internet-based", "computer-adaptive", "computer-delivered", "computer-aided English/language test", and "automatic scoring"—revealed that there are a good number of articles on computer-based English language testing published in Chinese academic journals. The vast majority of them, however, were think-pieces introducing, reviewing, or debating the challenges and potentials of computer-based English language tests and automated scoring systems; only a small number of publications reported empirical studies. Although these empirical studies are not typically of the depth or quality of Jin and Yan's or He and Min's (in this issue), this brief review of these studies sheds light on the current status of research on computer-based English tests in China. More importantly, the current review identifies research gaps and points to future research agendas, which can have implications beyond Chinese contexts to promote better use of computer technologies *in* and *for* language testing.

The Present

Three important research topics have appeared in publications on computer-based English language testing in China. The first has focused on computer-adaptive testing, driven to a great extent by the promising efficiency of test delivery, but constantly challenged by difficulties in designing appropriate techniques to assign the optimal number of items of appropriate difficulty levels to test takers. The second body of research, and arguably the largest in terms of number of publications, is related to the College English Test and achievement tests designed and used in individual universities via commercially available test systems. The third is concerned with the development of automated scoring systems to assess speaking, writing, and translation.

In terms of publication venues, *Foreign Language World [外语界]*, *Computer-Assisted Foreign Language Education [外语电化教学]*, and *Foreign Language Testing and Teaching [外语测试与教学]* (all published by Shanghai International Studies University) have been the three major journals publishing these empirical studies, followed by *Modern Foreign Language [现代外语]* (published by Guangdong University of Foreign Studies) and *Foreign Language Teaching and Research [外语教学与研究]* (published by Beijing Foreign Studies University). Other journals, such as the *China Information Processing Journal [中文信息学报]* and the *Journal of Tsinghua University [清华大学学报: 自然科学版]*, have mainly published studies on automated scoring systems.

Research on Computer-adaptive Testing

The earliest attempts to develop and research computer-based English tests in China started in Guangdong University of Foreign Studies in the 1990s and focused mainly on computer-adaptive testing. A number of articles were published by staff or former students of this

institution, for example, He (1999), Zhang (1999), Zeng (2002), and Huang and He (2013). He (1999) compared 55 university students' performance in a conventional paper-based test to their teachers' rankings of their language ability and to their performance on a "cognitive computer-adaptive test", which included reading comprehension, grammar and vocabulary, and cloze items. He (1999) found that the "cognitive computer-adaptive test" was more efficient, accurate, and consistent in assessing the participants' language abilities than was the conventional paper-based test. Zhang (1999) reported a high correlation (r=.86) between a computer-adaptive and a "self-adaptive" test which allowed the participants to decide the difficulty level of the next item presented to them. In Zhang's project, 50 university students completed multiple-choice vocabulary items. In addition to computer-adaptive and "self-adaptive," Zeng (2002) piloted "individualized self-adaptive testing" which required test takers not only to decide the difficulty level of the next item but also to report their confidence level in completing the test items (incorporated as a weighting element in scoring the participants' test performance). Zeng (2002) also reported that the addition of self-assessed confidence scores had some advantage over the use of maximum likelihood estimation in computer-adaptive testing.

Recently, Huang and He (2013) conducted Monte Carlo simulations to assess the usefulness of a three-parameter logistic graded model to address issues of local item dependence in a testlet in a computer-adaptive test of listening comprehension. Instead of statistical simulations, He and Min (this issue) used students' actual performance data in a conventional computer-based test and a corresponding computer-adaptive test, which included both dichotomously-scored items and polytomously-scored testlet-based items of reading and listening comprehension. Over 8,200 students took the conventional computer-based test, with 416 of them taking the computer-adaptive test, and completing a questionnaire designed to collect data such as the students' background profiles, their paper-based language test scores (in particular, on the CET), and indicators of computer familiarity. Shortly after taking the computer-adaptive test, some students voluntarily participated in focus-group discussions. He and Min reported that the conventional computer-based test and the computer-adaptive test were comparable and measured the same construct. Moreover, students' English language proficiency as measured by the paper-based test, unlike their computer familiarity, was a significant predictor of their performance in the two computer-based tests. Furthermore, factorial invariance of the computer-adaptive test scores of male and female students was noted, as another piece of supporting evidence of the quality of this computer-adaptive test, which is taken by undergraduates from Zhejiang University as part of their graduation requirement.

Research on the Computer-based CET and Institutional Achievement Tests at Universities

Du and Gui (2000) were among the first in China to develop their own system to explore the usefulness of computer-based English language tests alongside others cited immediately above. There were some sporadic efforts to develop computer-based English language tests[ii] in China between 2000 and 2005. However, it was the announcement in February 2005 (全

国大学英语四、六级考试改革方案<试行>) that the National CET Committee would consider using a computer-based CET that inspired, pushed, or influenced various stakeholders (e.g., teachers, researchers, universities, and publishers) to develop computer-based tests and testing systems. It was the commercial availability of a few computer-based language testing systems in the Chinese market (e.g., College English Oral Test System[iii] of Shanghai Foreign Language Education Press, iflytek[iv] as a spin-out publically listed company of the University of Science and Technology of China, Lange[v] of the Lancoo Group, and Wingsoft[vi] of Fudan University) that made it possible for universities and English language teachers to develop their own computer-based tests to assess their students on a large scale efficiently, because these systems often included a set of components for item development, delivery, marking, and data analysis and reporting.

Since 2005, a number of studies have reported local attempts to develop and use computer-based language tests in universities. These empirical studies explored a number of validity issues in the institutional computer-based achievement tests and the national IB-CET, such as the comparability between computer-based and paper-based tests in terms of students' performances on tests, the impact of computer literacy on test results and test-taking cognitive processes, the influence of delivery modes on the features of language produced in speaking and writing tasks, students' perceptions of and readiness for computer-based tests, and the advantages and the challenges of using computer-based tests. The majority of these studies have focused on the assessment of speaking and writing abilities, in other words, productive rather than receptive language abilities. This focus presents an interesting contrast with research on computer-adaptive tests which primarily used listening and reading tasks.

In what follows, we outline the key findings or research focuses of these studies in the chronological order of their publication. Qiu, Ji, Wan and Cheng (2005), using Wingsoft for test delivery at Fudan University, described their students' performances as well as the benefits and challenges in using computerised listening and speaking tasks such as read-aloud, summarization of pictures/cartoons, listening to news from foreign radio stations, and short debates between two test takers. Cai (2005), also using Wingsoft and at Fudan University, compared 182 students' performances in the computer-based and the face-to-face CET-Spoken English Test. He found a reasonable correlation (r= 0.71) between the two test modes. Similarly, Gao (2007) reported the comparability of students' performances on two computer-based speaking tests and one face-to-face speaking test at Hangzhou Dianzi University, observing generally positive attitudes of the students towards computer-based speaking tests of English. In a CET sponsored study, Cai and Wang (2009) compared computer-based and paper-based writing tests; they found that participants' typing speed, anxiety in typing, and computer familiarity did not significantly affect their writing performance. They also found that the marks assigned to the computer-processed and the hand-written scripts were highly correlated. From test takers' and teachers' perspectives, Li (2009) (in a study sponsored by the CET) investigated the extent to which the use of different types of speaking tasks (e.g., read-aloud, story completion, describing pictures, listening to retell/summarize, and group discussion) with multimedia input in end-of-term examinations might help to reduce test takers' anxiety differently. The speaking tasks in this

study were delivered via the Lange system. The key finding of Li's study was that test takers' anxiety was alleviated by the use of multimedia input – an existing practice of the IB-CET. Zhu and Zhang (2009) reported the benefits and challenges of using computerised listening and speaking tasks as part of the achievement tests at Ningbo University since 2006. Although the delivery platform was developed by their university's education technology company, Zhu and Zhang (2009) modelled their assessment tasks after Qiu et al. (2005) and reported similar findings. Tang and Liu (2009), from Beijing Foreign Studies University, reported that the performances of students with intermediate-level English proficiency were affected by the mode of test delivery, especially for the reading section of their computer-based test. Yin, Zheng, Wang and Xin (2010), from Harbin Institute of Technology, reported the differential effects of two delivery modes on 30 test takers' fluency, operationalized as seven features of speech such as average length of runs, pauses, errors, and subordinate clauses. Three types of tasks were used by Yin et al.: two non-interactive tasks (short-answer questions; answers to questions with supporting information provided) and one interactive task (triadic small group discussions on a given topic). They concluded that the students' fluency in non-interactive tasks was less likely to be affected by test delivery modes, while the interactive task (i.e., small group discussion) was more susceptible to test delivery modes, with significantly more errors and slower speaking rates observed in the computer-based speaking test.

Dai and You (2010) reported the results of Rasch analysis of six raters' severity and consistency in marking over 660 students' performances in three different types of computer-based speaking tasks; they pointed out that rater bias would not go away simply because of the use of computer-based testing. Dai (2011) then reported high comparability between face-to-face oral proficiency interviews (OPI) and computer-based OPI (COPI) and high consistency and similarity between two raters when marking OPI and COPI test performances. However, according to the data from a survey of test takers, the students preferred the OPI and considered COPI tasks less interactive than those in the OPI (see also Qian 2009 which reported similar findings based on data collected from Hong Kong university students). L. Jin (2011) reported on the practice of computer-based speaking tests in Inner Mongolia Normal University since 2005 via the Lange testing system. Xu, Xie, Liu, Chen, Liu and Gu (2013) reported a high correlation (r=0.91) between a teacher-administered face-to-face speaking test (involving a short-answer question and a topic-based dialogue between two test takers) and a computer-based speaking test (involving reading aloud a short passage and a topic-based monologue) delivered via iflytek. The students' performances on the computer-based test were automatically assessed through iflytek's automated evaluation system (see below). Based on data from the questionnaire survey given to the students and on interviews with six of their teachers, Xu et al. (2013) reported the generally positive attitudes of these stakeholders towards the computer-based speaking test.

The studies reviewed above all focused on English for general purposes. Research on computer-based assessment of English for specific purposes or content-based assessment has been rare in China, except for Si (2008) and Chen (2009). Si (2008) investigated the computer-based assessment of students' speaking ability in business contexts, while Chen (2009) piloted computer-based assessment of students' knowledge of English-Speaking

Countries. He found that the presentation of test materials in multimedia had a negative effect on the validity of the test.

As shown above, there have been several research studies on computer-based English language tests developed by individual universities; however, the nation-wide IB-CET, the largest computer-based English test in China, has not enjoyed the same extent of research effort. The first administration of the IB-CET in 56 universities (with 100 participants maximum from each institution) in December 2008 prompted a few studies designed to investigate the validity of the IB-CET and how well university students were coping with the new test formats. For example, Huang and Qin (2009) surveyed 36 test takers from a south-western university in the second week after the actual test. They found that the students were generally comfortable with the computer-based test formats, but probably needed more time to adjust themselves to the new listening and speaking tasks because the formats and time pressure of these tasks differed from the paper-based CET. Like Huang and Qin (2009), Liu (2011) surveyed 185 test takers for their views on (a) the key differences between the IB-CET and paper-based tests and (b) their challenges in different sections of the IB-CET. Similar to Huang and Qin (2009), Liu (2011) reported generally positive attitudes towards the IB-CET, but also suggested areas for improvement especially with regard to test content and computer interface. Drawing on data from questionnaire and interviews, Yang and Li (2010) reported a significant negative correlation between 52 test takers' computer anxiety and their perceptions of computer self-efficacy, as well as a significant positive correlation between computer anxiety and test anxiety in computer-based speaking tests. Jin and Yan (this issue) reported that test takers' high computer literacy can facilitate their performance in the IB-CET writing tasks (see also Jin & Wu, 2010) although their cognitive processes involved in the computer-based and paper-based writing tasks were similar. Jin and Yan argued that computer literacy should be considered as a contextual factor "closely related to the construct" being measured in computer-based tests, given the extensive use of computers in everyday life these days.

Research on Automated Scoring: Speaking, Writing and Translation

A limited amount of research in China has focused on three areas of automated scoring and test performance: speaking, writing, and translation. The scope of this article does not permit a discussion of natural language processing, computational linguistics, or statistical linguistics as part of a critique of automated scoring engines. Rather, we focus here on interpreting the reported correlations between the scores generated by the automated engines and the scores assigned by human raters, which is often claimed to be supporting evidence for the quality of automated scoring engines (cf. Chapelle & Douglas, 2006).

To the best of our knowledge, the IB-CET and PETS have implemented automated evaluation systems to assess speaking test performance (for Levels 1 and 2). However, there does not seem to be any published research on the automated evaluation system used in the IB-CET speaking test. For the PETS, Qiao, Dong and Liu (2012) described the components of the

automated scoring engine – EduRater. Based on about 1,000 test takers' performances in two different PETS-1 speaking tasks, which were marked by three raters independently, Qiao et al. reported a high correlation (r=0.81) between the automated and human scoring. Li, Yang, Feng, Wu, Chen and Hu (2008), Yan, Hu, Wei, Dai, Li, Yang and Feng (2009) and Yan, Hu, Wei, Li, Yang and Feng (2010) – a research team from the University of Science and Technology of China – reported on their attempts to develop an automated system[vii] to evaluate students' performance on speaking tasks – reading-aloud, retelling/summarization, and recitation, respectively. These three articles reported high correlations between automated and human scoring. However, Zhou and Zeng (2016) reported that automated and human scoring of high-school leavers' speaking performance differed significantly in terms of rater severity even though the overall distribution of students' test scores were similar across the two scoring methods. For the automated evaluation of writing performances, several researchers (e.g., Ge, 2010; Ge & Chen, 2009; Li & Ge, 2008; Li & Liu, 2013) have proposed slightly different models. Another interesting area of development is the automated evaluation of translations between English and Chinese (Jiang, 2013; Jiang & Wen, 2010, 2012; Liu & Liu, 2015; Wang & Chang, 2009; Wen, Qin & Jiang, 2009). These publications were based on doctoral dissertations, and all reported some kind of "superiority" for their automated scoring engines for evaluating speaking, writing, or translation; however, none of these engines has been externally validated beyond the initial thesis inquiries.

These three main research topics on computer-based English language testing demonstrate Chinese researchers' endeavours to better understand the efficiency of computer-based testing and the comparability between computer-based and paper-based tests alongside their efforts to develop automated scoring engines to evaluate speaking, writing, and translation performance. In addition, researchers have also been concerned about test takers' readiness for and attitudes towards computer-based tests as well as the fairness of computer-based tests for students of different experience and ability. These studies provide solid stepping stones into the future.

The Future

Computer technology is being used "in designing, developing and delivering test content as well as scoring and reporting examinee test performance" (Sawaki, 2012, p. 426). Chapelle (2010) listed three motives for using computer technology in language testing: efficiency, equivalence, and innovation. The use of computer technology can improve efficiency in test development, delivery and scoring, for example, via computer-adaptive testing and automated scoring of speaking and writing performances. Equivalence refers to the comparability in test performances between computer-based and paper-based or other traditional methods of assessment. By innovation, Chapelle suggested that the integration of technology can help to reconceptualize language ability as "the ability to select and deploy appropriate language through the technologies that are appropriate for a situation" (Chapelle & Douglas, 2006, p. 107). Douglas (2013, p. 2) urged that "we must define the language construct to include appropriate technology in light of the target situation and test purpose."

None of the empirical studies reviewed here has incorporated or viewed technology as part of the construct to be assessed in their tests (but see Jin & Yan, this issue, who call for a reconceptualization of the construct of computer-based writing tests). These past studies have simply been used or considered computer as a tool rather than as an integral aspect of the construct being assessed. Given the extensive use of computer technology in language learning and communication nowadays, especially in higher education and work, it is high time to reconceptualize what is meant by computer-based language testing. In our view, such a reconceptualization should be the premise and the guiding rationale for any innovation in assessment practices. As Chalhoub-Deville (2010, p. 522) contended, "L2 CBTs, as currently conceived, fall short in providing any radical transformation of assessment practices." It is evident, in the studies reviewed above as well as in other publications in international academic journals, that various terms are used to refer to computer-based language testing, including: computer-adaptive, computer-aided, computer-assisted, computer-enhanced, computer-mediated, computer-supported, and technology-enhanced. All these terms imply that computers play a peripheral role, as a supporter, enhancer, or mediator of communication and the demonstration of language abilities. For innovations in computer-based language testing to really occur, it is imperative that researchers, assessors, and educators consider computers not only as a delivery platform but also as an integrated part of the language construct to be assessed. Following this notion, hereafter we recommend using the term computer-integrated[viii] instead of computer-based. However, there is a sensitive balance to strike. As Milanovic (2013, p. 32) put it, "we must try to take advantage of the benefits technology has to offer without the technology tail wagging the learning and assessment dog." Or, in Douglas' (2000, p. 275) words, "language testing…driven by technology, rather than technology being employed in the services of language testing, is likely to lead us down a road best not traveled." Douglas (2013, p. 6) further reminded us that "the use of technology for its own sake can lead to the trivializing of language test tasks by limiting what we can include in our tests to those things that can be delivered easily by computers or the Internet or that can be scored easily by machines."

The reconceptualization of computer-integrated language testing will facilitate and promote innovations in test design, especially in task formats and assessment criteria. As a result, there needs to be a shift of research focus. There should be fewer studies on the comparability between computer-integrated and paper-based tests or on the adverse or beneficial impacts of computer literacy on test performance, as in the vast majority of the previous studies in China. Instead, more studies need to focus on the comparability between different computer technologies or platforms. If computer literacy is accepted as an essential part of the language construct to be assessed, computer-integrated and paper-based tests may no longer share the same level of comparability, in theory and by design, as current computer-based and paper-based tests do. Instead, future comparability studies may focus on comparability between different computer-delivery platforms, posing research questions such as, How does language performance differ according to the use of certain technologies? What has been considered as critical features of test delivery environments and test takers' computer literacy in the past or at present may soon become outdated or irrelevant. For example, the screen resolution of desktops as researched in Bridgeman, Lennon and Jackenthal (2003) a decade ago is hardly a controversial issue now.

Computer literacy is an evolving concept, which must necessarily be a long-lasting concern of test providers, given the rapid development of different computer technologies and platforms (desktops, laptops, tablets, virtual worlds, etc.). For example, studies in the 1990s (Powers & Potenza, 1996) and 2010s (Ling, 2016; Ling & Bridgeman, 2013) on the comparability between writing on desktop and laptop computers produced conflicting findings. In Powers and Potenza (1996), participants were in favour of desktop computers, and essays written on desktop computers achieved higher scores than those written on laptop computers. However, in Ling and Bridgeman (2013, p.118), "essays produced using a laptop were comparable to those produced using desktop computers on essays' scores, lengths, and writing speed." Ling (2016) reported that taking a test on an iPad was similar to taking a test on a desktop computer for experienced users of these two types of devices. Future computer technologies may become more interactive and intuitive to use. Multimodal delivery of language test content (visual, audio, animations, virtual reality, etc.) via computer technology, which was not possible in paper-based tests, may better represent the progressively evolving construct of language use and hence improve task authenticity. In this respect, it will be fruitful to promote efforts to research innovative multimodal and interactive tasks, the inter-operationability of such tasks in different platforms, and their differential impacts on performance of test takers of different characteristics and for different assessment purposes.

It may, however, take a long time and considerable effort to reconceptualize the construct of computer-integrated language testing and to shift the focus of research from comparability to issues in multimodality and inter-operationability. At least three other areas require immediate action and can produce more fruitful research evidence to broaden the scope, depth, and quality of research on the current practices of computer-integrated English language testing in China.

Firstly, given the stakes and the impacts that the IB-CET has on teaching, learning and Chinese society in general, more high quality research studies on the IB-CET, whether independent from or commissioned by the CET, are urgently needed. The IB-CET is the largest computer-based high-stakes English test in China, however, there are only a small number of research publications on the IB-CET in Chinese academic journals or elsewhere; these studies tended to be small-scale and conducted prior to 2011. There does not seem to be any research publication on the automated evaluation system used for marking the IB-CET speaking task performance. This lack may be because the IB-CET is a live, consequential test and therefore for security reasons data about the test or its tasks are not released beyond the National CET Committee. Nevertheless, there should be more publications evaluating the IB-CET, such as Jin and Yan's (this issue), if the Committee can release data for research purposes.

Secondly, studies on the comparability of cognitive processes among examinees taking computer-based and paper-based tests would be a welcome addition to current knowledge about the effects of the two delivery modes. Almost all the Chinese empirical studies on the comparability between computer-based and paper-based tests, or the impacts of computer literacy on test performance, have relied primarily on test results as research data, except for Jin and Yan's (this issue) investigations of test-taking processes. More research is needed

to investigate the comparability between the two delivery modes in terms of test-taking cognitive processes at an individual as well as at group-level (Yu, 2010). Taking computer literacy as an example, group-level analyses might have demonstrated that students' test scores were not affected by their computer literacy, however, at individual level, computer literacy might well affect certain test taker's cognitive process (Yu, 2010). Computer technology provides ample opportunities and data to do this kind of research. Test takers' response time, keyboarding speed, confidence level, and test taking efforts (Setzer, Wise, van den Heuvel, & Ling, 2013), just to name a few sources of data, can be readily recorded. Streamlining computer-based language tests with eye-tracking devices provides further opportunities to record students' eye movements as an indicator of their attentional and test-taking processes (see Yu, He & Isaacs, in press). Analyses of test takers' cognitive processes can help not only to understand the validity of the tasks but also to deter and detect cheating or task-irrelevant behaviours during a test. Preventing cheating and enhancing test security have been one of the motives for creating the current IB-CET (see Jin & Yan, this issue).

Thirdly, more independent, transparent, and comparative research on the quality of automated evaluation engines is needed to assure test takers that they are assessed fairly. All the Chinese empirical studies to date have reported how well their automated evaluation systems predicted human scores; however, as Carr (2014) rightly pointed out, research on automated evaluation systems "has been conducted by the companies developing the systems, and…there is a marked lack of independent research comparing different systems head to head." This limitation also applies to the existing Chinese studies. Independent research is needed to advance technological breakthroughs as well as transparency. In addition to more rigorous and independent research, it is equally important to expand the focus of research on automated scoring. There are a number of high priority topics that have hardly been explored, such as how test takers interact with tasks that use automated scoring, what test-taking processes and strategies appear, how score users (e.g., university admission tutors and language support staff) interpret and use the test scores assigned by automated scoring engines, and the impact of the use of automated scoring on language teaching, learning, and test preparation. Future automated evaluation engines should build not only on what computers can do but also on what the construct of computer-integrated communication or language performance should be.

Conclusion

This article provides a snapshot of the current research and practice on computer-integrated English language testing in China as conducted by Chinese scholars and published in major Chinese academic journals. The quality of these publications is not at the same level as appears in Jin and Yan (this issue) and He and Min (this issue); nevertheless, the themes and focuses of these studies help to identify three key areas of current endeavours in researching and using computer-based English language testing in China: computer-adaptive testing, the national IB-CET and certain institutional achievement tests at universities, and automated evaluation systems for speaking, writing, and translation assessment.

Considering these research focuses and findings, two areas should shape future agendas for research. Firstly, the construct of language use in computer-based tests needs reconceptualizing by integrating computer technology not only as a delivery platform but also as an integral component of the language construct to be assessed. This innovation should facilitate and promote innovations in test design, especially in task formats and assessment criteria. Secondly, as a consequence, comparability studies should shift their research focuses from studying the comparability between computer-based and paper-based tests to studying the comparability between different computer technologies and platforms, the inter-operationability of innovative multimodal and interactive tasks in different delivery platforms, and their potential impact on the performance of test takers of different characteristics for different assessment purposes. However, given the current status of research on computer-based English language testing in China three more pressing issues require immediate action: (a) more high quality research on the validity of the IB-CET and its automated evaluation of speaking task performance, (b) more high quality research on students' test taking cognitive processes, and (c) more independent, transparent, and comparative research on the quality of automated evaluation engines, which should be based on the construct of computer-integrated testing, rather than on the construct of traditional paper-based testing or communication. The findings from these studies will have implications beyond Chinese contexts in promoting better use of computer technologies in and for language assessment.

Acknowledgements

The Editors of this special issue—Alister Cumming and David Qian—and Professor Lianzhen He of Zhejiang University during her visit as Benjamin Meaker Visiting Professor at the University of Bristol provided insightful comments and feedback on an earlier draft of this article. This article was first presented as a keynote speech at the inaugural conference of the Asian Association of Language Assessment in October 2014 in Hangzhou, China.

References

Bridgeman, B., Lennon, M. L., & Jackenthal, A. (2003). Effects of screen size, screen resolution, and display rate on computer-based test performance. *Applied Measurement in Education, 16*(3), 191-205.

Cai, J. (2005). Validity, reliability and practicality of computer-based oral proficiency test. *Foreign Language World*, *26*(4), 66-75. [蔡基刚. (2005). 大学英语四, 六级计算机口语测试效度, 信度和可操作性研究. *外语界, 26*(4), 66-75.]

Cai, J. & Wang, Z. (2009). A study of validity and reliability of Internet-based English writing testing *Foreign Language World*, *30*(3), 52-58. [蔡基刚, & 汪中平. (2009). 英语网考的写作效度和信度研究. *外语界, 30*(3), 52-58.]

Carr, N. T. (2014). Computer-automated scoring of written responses. In A. Kunnan (Ed.), *The Companion to language assessment*. Malden, MA: John Wiley & Sons, Inc. DOI: 10.1002/9781118411360.wbcla124

Chalhoub-Deville, M. B. (2010). Technology in standardized language assessments. In R. Kaplan (Ed.), *The Oxford handbook of applied linguistics* (2nd ed.) (pp. 511-526). Oxford: Oxford University Press.

Chapelle, C. A. (2010). Technology in language testing. In Fulcher, G. & Trasher, R. *Language testing videos*. In association with ILTA. Available: http://languagetesting.info.

Chapelle, C. A., & Douglas, D. (2006). *Assessing language through computer technology*. Cambridge, UK: Cambridge University Press.

Chen, H. (2009). A proposal on the verification model of the validity equivalence between PBLT and CBLT. *Foreign Language World*, *30*(3), 73-80 [陈慧麟. (2009). 基于纸笔的语言测试和基于计算机的语言测试之间效度对等性验证模式初探. *外语界, 30*(3), 73-80.]

Cheng, L. (2008). The key to success: English language testing in China. *Language Testing, 25*(1), 15-37.

Dai, Z. (2011). A study of the reliability of computerized oral proficiency interview. *Computer-Assisted Foreign Language Education*, *33*(2), 45-50. [戴朝晖. (2011). 计算机口语考试信度研究. *外语电化教学, 33*(2), 45-50.]

Dai, Z. & You, Q. (2010). Multi-facets Rasch Model analysis of rater bias in Computerized EFL Oral Proficiency Interview. *Foreign Language World*, *31*(5), 87-95. [戴朝晖, & 尤其达. (2010). 大学英语计算机口语考试评分者偏差分析. *外语界, 31*(5), 87-95.]

Douglas, D. (2000). *Assessing language for specific purposes*. Cambridge, UK: Cambridge University Press.

Douglas, D. (2013). Technology and language testing. In C. A. Chapelle (Ed.), *The Encyclopedia of Applied Linguistics*. Malden, MA: Wiley-Blackwell. DOI: 10.1002/9781405198431.wbeal1182

Du, J. & Gui, S. (2000). An experimental study of computerized diagnostic testing of reading. *Foreign Language Teaching and Research*, *32*(5), 345-351. [杜金榜, & 桂诗春. (2000). 电脑化阅读诊断测试的实验研究. *外语教学与研究, 32*(5), 345-351.]

Gao, B. (2007). A comparative study of COPT and DOPT. *Computer-Assisted Foreign Language Education*, *29*(2), 73-76. [高丙梁. (2007). 计算机口试与面试的比较研究. *外语电化教学, 29*(2), 73-76.]

Ge, S. (2010). A comparative study of automated essay scoring techniques for college students' English writing. *Journal of Guangdong University of Foreign Studies*, *21*(3), 87-90. [葛诗利. (2010). 大学英语作文自动评分方法比较研究. *广东外语外贸大学学报, 21*(3), 87-90.]

Ge, S. & Chen, X. (2009). Cluster analysis of college English writing in automated essay scoring. *Computer Engineering and Applications*, *45*(6), 145-148. [葛诗利, & 陈潇潇. (2009). 文本聚类在大学英语作文自动评分中应用. *计算机工程与应用, 45*(6), 145-148.]

He, L. (1999). Designing cognitive computer-adaptive tests. *Modern Foreign Languages*, *22*(2), 169-183. [何莲珍. (1999). 认知计算机适应性考试模型的设计. *现代外语, 22*(2), 169-183.]

Huang, M. & Qin, C. (2009). An investigation into test takers' adaptability to the Internet-based CET. *Foreign Language World*, *30*(5), 90-96. [黄敏, & 覃朝宪. (2009). 全国大学英语网络考试考生情况调查——以适应性为研究视角. *外语界, 30*(5), 90-96.]

Huang, Y. & He, L. (2013). Approach to fitting testlet for computerized adaptive language testing. *Computer-Assisted Foreign Language Education*, *35*(2), 29-34. [黄妍, & 何莲珍. (2013). 计算机自适应语言测试的题组拟合方法. *外语电化教学, 35*(2), 29-34.]

Jiang, J. (2013). An automatic approach to evaluating the linguistic quality of English-Chinese translations. *Modern Foreign Languages*, *36*(1), 85-91. [江进林. (2013). 英译汉语言质量自动量化研究. *现代外语, 36*(1), 85-91.]

Jiang, J. & Wen, Q. (2010). A comparative study of N-gram and translation unit alignment in automated scoring of students' English-Chinese translation. *Modern Foreign Languages*, *33*(2), 177-184. [江进林, & 文秋芳. (2010). N 元组和翻译单位对英译汉自动评分作用的比较研究. *现代外语, 33*(2), 177-184.]

Jiang, J. & Wen, Q. (2012). Computer scoring models for EFL learners' English-Chinese translation in large-scale tests. *Computer-Assisted Foreign Language Education*, *34*(2), 3-8. [江进林, & 文秋芳. (2012). 大规模测试中学生英译汉机器评分模型的构建. *外语电化教学, 34*(2), 3-8.]

Jin, L. (2011). A brief study of computer-aided test of oral English. *Foreign Language and Literature*, *27*(4), 126-130. [金力. (2011). 计算机辅助大学英语口语测试研究. *外国语文, 27*(4), 126-130.]

Jin, Y., & Wu, J. (2010). A preliminary study of the validity of the Internet-based CET-4—Factors affecting test-takers' perception of and performance on the test. *Computer-Assisted Foreign Language Education*, *32*(2), 3-10. [金艳, & 吴江. (2010). 大学英语四级网考效度初探——影响考生评价和考试成绩的因素分析. *外语电化教学, 32*(2), 3-10.]

Li, M., Yang, X., Feng, G., Wu, M., Chen, J., & Hu, G. (2008). Machine scoring of reading aloud item of large-scale college English oral tests. *Foreign Language World*, *29*(4), 88-95. [李萌涛, 杨晓果, 冯国栋, 吴敏, 陈纪梁, & 胡国平. (2008). 大规模大学英语口语测试朗读题型机器阅卷可行性研究与实践. *外语界, 29*(4), 88-95.]

Li, X. & Liu, J. (2013). Ensemble learning based essay automated scoring algorithm for Chinese English learners. *Journal of Chinese Information Processing*, *27*(5), 100-106. [李霞, & 刘建达. (2013). 适用于中国外语学习者的英文作文全自动集成评分算法. *中文信息学报, 27*(5), 100-106.]

Li, Y. (2009). An empirical study of the effect of the large-scale computer-assisted Spoken English Test. *Foreign Language World*, *30*(4), 69-76. [李玉平. (2009). 大规模计算机辅助英语口语测试效果实证研究. *外语界, 30*(4), 69-76.]

Li, Y. & Ge, S. (2008). The validity of word list in automated essay scoring for college students. *Foreign Languages and Their Teaching*, *24*(10), 48-52. [李艳, & 葛诗利. (2008). 大学英语作文自动评分中分级词表的效度研究. *外语与外语教学, 24*(10), 48-52.]

Ling, G. (2016). Does it matter whether one takes a test on an ipad or a desktop computer? *International Journal of Testing*, 16(4), 352-377. doi: 10.1080/15305058.2016.1160097

Ling, G., & Bridgeman, B. (2013). Writing essays on a laptop or a desktop computer: Does it matter? *International Journal of Testing, 13*(2), 105-122. DOI: 10.1080/15305058.2012.690012

Liu, P. (2011). Investigation on the problems of Internet-based CET 4 & 6 and solutions from test-takers' perspective. *Modern Educational Technology*, *21*(12), 77-81. [刘萍. (2011). 考生视角下大学英语四, 六级网考的问题与对策. *现代教育技术, 21*(12), 77-81.]

Liu, Z. & Liu, D. (2015). The design and implementation of a system for the automatic assessment of learners' translations. *Journal of PLA University of Foreign Languages*, *38*(2), 109-115.[刘泽权. &刘鼎甲. (2015). 学习者英译文自动评估系统的设计与实现. 解放军外国语学院学报, *38*(2), 109-115.]

Milanovic, M. (2013). A look into the future. *Research Notes* (51), 31-33.

Powers, D. E., & Potenza, M. (1996). *Comparability of testing using laptop and desktop computers* (ETS RR-96-15). Princeton, NJ: Educational Testing Service.

Qian, D. D. (2009). Comparing direct and semi-direct modes for speaking assessment: Affective effects on test takers. *Language Assessment Quarterly, 6*(2), 1-13.

Qiao, H., Dong, B., & Liu, C. (2012). A study on automated scoring of PETS computer-based speaking test. *Foreign Language Testing and Teaching*, *2*(3), 47-52. [乔辉, 董滨, & 刘常亮. (2012). PETS 计算机辅助口试自动评分技术研究. *外语测试与教学, 2*(3), 47-52.]

Qiu, D., Ji, P., Wan, J. & Cheng, Y. (2005). A study on computer-based listening and speaking tests for college students. *Foreign Language World*, *26*(4), 76-79 . [邱东林, 季佩英, 万江波, & 程寅. (2005). 大学英语听说机考尝试. *外语界, 26*(4), 76-79.]

Sawaki, Y. (2012). Technology in language testing. In G. Fulcher & F. Davidson (Eds.), *The Routledge handbook of language testing* (pp. 426-437). Abingdon, UK: Routledge.

Setzer, J. C., Wise, S. L., van den Heuvel, J. R., & Ling, G. (2013). An Investigation of examinee test-taking effort on a large-scale assessment. *Applied Measurement in Education, 26*(1), 34-49. doi: 10.1080/08957347.2013.739453

Si, Y. (2008). An empirical study of large-scaled computer-assisted diagnostic testing of spoken business English. *Computer-Assisted Foreign Language Education*, *30*(1), 67-71 . [司耀龙. (2008). 基于计算机的大规模商务英语口语诊断测试实践研究. *外语电化教学, 30*(1), 67-71.]

Tang, J. & Liu, X. (2009). Effects of delivery mode on test performance – A comparative study on computer-based and paper-based tests. *Distance Education in China*.唐锦兰, & 刘晓悦. (2009). 考试媒介对于考生成绩的影响研究——一项英语机考与纸笔考试成绩对比分析. *中国远程教育*(5), 57-61.

Wang, L. & Chang, B. (2009). Research on the human-aided auto-assessment for translation tests in College English. *Computer-Assisted Foreign Language Education*, *31*(4), 17-21. [王雷, & 常宝宝. (2009). 大学英语翻译考试人工辅助计算机评分初探. *外语电化教学, 31*(4), 17-21.]

Wang, Y. (2004). A study of online marking of CET compositions. *Foreign Language World*, *25*(5), 74-79. [王跃武. (2004). 大学英语四, 六级考试作文网上阅卷实验研究. *外语界, 25*(5), 74-79.]

Wang, Y., Zhu, Z., Yang, H. (2006). The implementation of a many-facet Rasch measurement to the reliability estimates of online marking *Foreign Language World*, *27*(1), 69-76 . [王跃武, 朱正才, & 杨惠中. (2006). 作文网上评分信度的多面 Rasch 测量分析. *外语界, 27*(1), 69-76.]

Wen, Q. , Qin, Y., & Jiang, J. (2009). Application of bilingual alignment technology to automatic translation scoring of English test. *Computer-Assisted Foreign Language Education*, *31*(1), 3-8 . [文秋芳, 秦颖, & 江进林. (2009). 英语考试翻译自动评分中双语对齐技术的应用. *外语电化教学, 31*(1), 3-8.]

Xu, Z. , Xie, X., Liu, C. , Chen, X., Liu, F., & Gu, J. (2013). An empirical study on large-scale computer-assisted college oral English test. *Modern Educational Technology*, *23*(8), 76-80. [徐智鑫, 谢小苑, 刘长江, 陈向俊, 刘芳, & 谷健飞. (2013). 高校大规模计算机辅助英语口语测试实证研究. *现代教育技术, 23*(8), 76-80.]

Yan, K., Hu, G., Wei, S., Dai, L. , Li, M., Yang, X., & Feng, G. (2009). Automatic evaluation of English retelling proficiency in large machine based oral English tests. *Journal of Tsinghua University (Science and Technology)*, *49*(S1), 1356-1362. [严可, 胡国平, 魏思, 戴礼荣, 李萌涛, 杨晓果, & 冯国栋 (2009). 面向大规模英语口语机考的复述题自动评分技术研究. *清华大学学报: 自然科学版, 49*(S1), 1356-1362.]

Yan, K., Hu, G., Wei, S., Li, M., Yang, X. & Feng, G. (2010). A primary study on computerised automatic marking of English recitation proficiency. *Computer Applications and Software*, *27* (7), 164-168. [严可, 胡国平, 魏思, 李萌涛, 杨晓果, & 冯国栋 (2010). 计算机用于英语背诵题的自动评分技术初探. *计算机应用与软件. 27* (7), 164-168.]

Yang, Y. & Li, M. (2010). A study on attitudes of college students in the computer-based oral English test environment. *Foreign Language World, 31*(6), 78-84. [杨艳霞, & 李萌涛. (2010). 大学英语口语机考环境下大学生计算机使用态度研究. *外语界, 31*(6), 78-84.]

Yin, N., Zheng, Y., Wang, L. , & Xin, D. (2010). A Comparative study on the effects of COPT and OPI on oral fluency. *Computer-Assisted Foreign Language Education, 26*(3), 25-29. [尹楠, 郑玉荣, 王丽丽, & 辛丹. (2010). 机辅与面试对口语流利性影响的对比研究. *外语与外语教学, 26*(3), 25-29.]

Yu, G. (2010). Effects of presentation mode and computer familiarity on summarization of extended texts. *Language Assessment Quarterly, 7*(2), 119-136.

Yu, G., He, L., & Isaacs, T. (in press). *The cognitive processes of taking IELTS Academic writing task one: An eye-tracking study*. Report to British Council/Cambridge Assessment.

Zeng, Y. (2002). A preliminary study on individualized self-adaptive testing. *Foreign Language Teaching and Research*, *34*(4), 278-282. [曾用强. (2002). 个性化自适应性测试探索. *外语教学与研究, 34*(4), 278-282.]

Zeng, Y. (2010). The Computerized Oral English Test of the National Matriculation English Test. In L. Cheng & A. Curtis (Eds.), *English language assessment and the Chinese learner* (pp. 234-247). Abingdon, UK: Routledge.

Zhang, S., & Yu, P. (2010). Online rater training for CET4 writing assessment. *Foreign Language World*, *31*(5), 79-86. [张森, & 于朋. (2010). 大学英语四级考试作文网上评阅信度保障研究. *外语界, 31*(5), 79-86.]

Zhang, W. (1999). An experiement with self-adaptive testing. *Journal of PLA University of Foreign Languages*, *22*(3), 53-55. [张武保. (1999). 自适应性测试的实验研究. *解放军外国语学院学报, 22*(3), 53-55.]

Zhou, Y. & Zeng, Y. (2016). Many-facet Rasch model analysis of computer automatic scoring in a computer-based English listening-speaking test. *Foreign Language Testing and Teaching*, *6*(1), 22-31.[周燕, &曾用强. (2016). 机助英语听说考试计算机自动评分的多层面 Rasch 模型分析. *外语测试与教学. 6*(1), 22-31.]

Zhu, Y. & Zhang, X. (2009). An exploration of practice on the computer-based testing in College English. *Computer-Assisted Foreign Language Education*, *31*(2), 63-67. [朱音尔, & 张肖莹. (2009). 基于网络的大学英语机考探索与实践. *外语电化教学, 31*(2), 63-67.]

Notes

[i] www.cnki.net; the database houses full-texts of all current Chinese journals (from 1915 onwards).

[ii] Before the CET was computerized, efforts to use computer technology in CET mainly focused on "online" marking of writing for the paper-based CET in order to improve and monitor marking reliability and efficiency. The writings were scanned to be marked. A few CET-sponsored studies reported the benefits of using online marking over "conference marking" thanks to the real-time monitoring function of the online marking system (Wang, 2004; Wang, Zhu & Yang, 2006; Zhang & Yu, 2010). Although strictly speaking, these studies were not about computer-based tests, the current practice of online marking of writings produced in computer-based CET has been influenced by the findings of these studies, and therefore, we think they are worth mentioning here.

[iii] ISBN of the product: 978-7-900717-85-6/H·53; Listed Price: 200k Chinese Yuan; http://www.ssit.cc/product_in.aspx?PID=24&CategoryName=%E5%A4%A7%E5%AD%A6%E8%8B%B1%E8%AF%AD&CID=1; The Test System was developed in collaboration with the University of Science and Technology of China.

[iv] http://www.iflytek.com/en/index.html. The system has an automated evaluation component.

[v] http://www.gzlange.com/paperless-examination.aspx

[vi] http://www.wingsoft.com.cn/product3.jsp

[vii] This university developed iflytek which contains automated evaluation of speaking performances, so it is possible that it was these researchers who developed the automated evaluation system in iflytek.

[viii] However, this proposal is not to suggest that all of these terms should be replaced by just one term, i.e., computer-integrated assessment, as each term tends to have specific meanings and operate in different assessment contexts. In cases where the use of computer technology is deemed to be part of the construct of the language task/test, it is more appropriate to use the term "computer-integrated" assessment.